UNITED STATES PATENT APPLICATION FOR

METHOD AND SYSTEM FOR COMPENSATING FOR PARALLAX IN MULTIPLE CAMERA SYSTEMS

INVENTOR:

Jonathan T. Foote

CERTIFICATE OF MAILING BY "EXPRESS MAIL" UNDER 37 C.F.R. § 1.10

"Express Mail" mailing label number: **EL 622 697 704 US**
Date of Mailing: November 20, 2001

I hereby certify that this correspondence is being deposited with the United States Postal Service, utilizing the "Express Mail Post Office to Addressee" service addressed to **Box PATENT APPLICATION, U.S. Patent and Trademark Office, P.O. Box 2327, Arlington, VA 22202** and mailed on the above Date of Mailing with the above "Express Mail" mailing label number.

Austin Gloria
Signature Date: November 20, 2001

METHOD AND SYSTEM FOR COMPENSATING FOR PARALLAX IN MULTIPLE CAMERA SYSTEMS

Inventor:

Jonathan T. Foote

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application is a continuation-in-part of Application Serial No. 09/370,406 entitled "AUTOMATIC VIDEO SYSTEM USING MULTIPLE CAMERAS" filed August 9, 1999, inventors Jonathan T. Foote, Subutai Ahmad and John Boreczky.

COPYRIGHT NOTICE

[0002] A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

FIELD OF THE INVENTION

[0003] This invention relates to video systems and composition of multiple digital images. The invention is more particularly related to the composition of multiple digital images each captured by an individual camera of a camera array,

and to a method and system for warping and combining at least portions of images so that a single image can be produced without seams or object disparity.

BACKGROUND

[0004] Remote and locally located cameras typically include devices for camera control. Devices include stepping motors or other mechanisms configured to point the camera or an image capturing device toward a scene or point of interest. Examples include teleconferencing applications, surveillance cameras, security cameras, cameras that are remotely controlled, activated by motion, light or other stimuli, remote sensing cameras such as those placed on robotic means (examples including those used in space exploration, deep sea diving, and for sensing areas or scenes to dangerous or inaccessible for normal camera operations (inside nuclear reactor cores, inside pipes, police cars, or law enforcement robotics, for example).

[0005] Normally, cameras are manually operated by a human operator on site, or remotely controlling the camera via a steering input (joystick or mouse, for example). In the case of remotely steered cameras, steering inputs generally activate a control program that sends commands to a stepping motor or other control device to steer a camera toward an object, item, or area of interest. General zooming functions of the camera may also be activated either on site or remotely.

[0006] In the case of teleconferencing applications (meetings, lectures, etc.), a variable angle camera with a mechanical tilt, pan, focal length, and zoom capability is normally used. Such devices generally require a human operator to

orient, zoom, and focus a video or motion picture camera. In some cases, conference participants may be required to activate a specific camera or signal attention of a camera configured to zoom in or focus on selected areas of a conference room.

[0007] Multiple cameras have been utilized in a number of applications. For example, Braun et al., U.S. Patent No. 5,187,571, "TELEVISION SYSTEM FOR DISPLAYING MULTIPLE VIEWS OF A REMOTE LOCATION," teaches an NTSC camera array arranged to form an aggregate field, and Henley, U.S. Patent No. 5,657,073, "SEAMLESS MULTI-CAMERA PANORAMIC IMAGING WITH DISTORTION CORRECTION AND A SELECTABLE FIELD OF VIEW," teaches a system for production of panoramic/panospheric output images.

[0008] Applications for multiple or steerable cameras include teleconferencing systems that typically direct a camera toward a speaker who is then broadcast to other teleconference participants. Direction of the camera(s) can be performed manually, or may utilize a tracking mechanism to determine a steering direction. Some known tracking mechanisms include, Wang et al., "A Hybrid Real-Time Face Tracking System," in Proc. ICASSP '98, and, Chu, "Superdirective Microphone Array for a Set-Top Videoconferencing System," In Proc. ICASSP '97.

[0009] However, technical challenges and costs have prevented such systems from becoming common and in wide spread use.

[0010] Systems attempting to integrate multiple images have failed to meet the needs or goals of users. For example, McCutchen, U.S. Patent No. 5,703,604,

"IMMERSIVE DODECAHEDRAL VIDEO VIEWING SYSTEM," teaches an array of video cameras arrayed in a dodecahedron for a complete spherical field of view. Images are composed at the receiving end by using multiple projectors on a hemispherical or spherical dome. However, the approach taught in McCutchen will suffer problems at image boundaries, as the multiple images will not register perfectly and result in obvious "seams." Additionally, object disparity may result due to combining overlapping images taken from multiple cameras.

[0011] In another example, Henley et al., U.S. Patent No. 5,657,073, "SEAMLESS MULTI-CAMERA PANORAMIC IMAGING WITH DISTORTION CORRECTION AND SELECTABLE FIELD OF VIEW," teaches combining images from radially-arranged cameras. However, Henley fails to disclose any but radially-arranged cameras, and does not provide details on image composition methods.

[0012] In another example, Nalwa, U.S. Patent No. 5,745,305 issued on April 28, 1998, entitled "PANORAMIC VIEWING APPARATUS, " teaches a four-sided pyramid-shaped element to reflect images from four different directions to four different cameras. Each camera is positioned to receive a reflected image from one of the reflective sides of the pyramids. The cameras are arranged so that they share a perceived Center Of Projection ("COP") located at the center of the pyramid. A common COP means there can be no overlap region to blend camera images, so seams will be more apparent.

[0013] Panoramic viewing systems using optical means to achieve a common COP are expensive and difficult to manufacture to the necessary degree

of precision. Manufacturing requires precision alignment of the pyramid and cameras to generate the perceived COP. If one of the cameras is initially misaligned, subsequently moved or the pyramid element is shifted, distortion may occur between combined images.

[0014] Accordingly, there is a desire to provide a cheaper, more reliable, and faster multiple camera viewing system for generating panoramic images.

SUMMARY

[0015] Roughly described, the present invention utilizes a camera array to capture plural piecewise continuous images of a scene. Each of the images are warped to a common coordinate system and overlapping images are combined using a technique to reduce parallax, thereby producing a single, seamless image of the scene.

[0016] According to an aspect of the invention, cameras are registered such that the processing equipment used to combine the images knows which parts of each image are overlapping or have the same points of the scene captured such that they may be precisely combined into a single image or scene.

[0017] Each captured image is warped to a common coordinate system and points of a scene captured by two or more cameras are combined into a single point or set of points appropriately positioned according to the scene being captured. Disparity correction may be used to combine overlapping images. Blending techniques may also be applied to edges of each of the images to remove any brightness or contrast differences between the combined images.

[0018] In an example, the present invention may be embodied in a device, or a camera array, comprising, a set of cameras mounted in an array, an image combining mechanism configured to combine at least two images captured from said set of cameras into a composite image, a view selection device configured to select a view from the composite image, and an output mechanism configured to display the selected view.

[0019] Another aspect of the invention may also include a method of registering a camera array, comprising the steps of, placing at least one registration point in a field of view of at least two cameras of said camera array, identifying a location of each registration point in a field of view of each camera of said array, and maintaining information about each registration point in relation to each camera such that images may be combined in relation to said registration points.

[0020] According to another aspect, a method for combining at least a portion of a plurality of images is provided. The method obtains a first image and a second image, wherein at least a portion of said first image and said second image include a common field of view. After obtaining the images, the method adjusts at least a portion of one of the images to reduce image disparity between the common field of view of the images and combines at least a portion of the first and second images to form a composite image.

[0021] In still another aspect, a method for combining frames of video from a plurality of cameras arranged in a camera array is provided. The method includes the steps of obtaining a first frame from a first camera, and a second frame from a second camera, wherein at least a portion of said first frame and said second

frame include a common field of view, stretching at least a portion of said first frame to reduce an image disparity between said common field of view of said first frame and said second frame and, combining said common field of view of said first frame and said second frame subsequent to said step of stretching.

[0022] According to another aspect, a method for combining a plurality of images captured from a plurality of cameras of a camera array into a panoramic image is provided. The method includes the steps of adjusting a first portion of a first image to reduce image disparity between said first portion of said first image and a second image, adjusting a second portion of said first image to reduce image disparity between said second portion of said first image and a third image; and, combining said first image, said second image, and said third image into a panoramic image.

[0023] According to still another aspect, an apparatus for producing a panoramic video is provided. The apparatus comprises a camera array including a plurality of cameras and an image obtaining device. The image obtaining device is configured to obtain images from the cameras in the camera array which share a common field of view. The apparatus also include an image adjustor which is configured to adjust at least a portion of said first image to reduce an image disparity between said common field of view of said first image and said second image. An image combiner is also included in the apparatus which combines at least a portion of images after adjusting.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] A more complete appreciation of the invention and many of the attendant advantages thereof will be readily obtained as the same becomes better understood by reference to the following detailed description when considered in connection with the accompanying drawings, wherein:

[0025] Fig. 1A is an example of a circular array of video cameras, according to the present invention;

[0026] Fig. 1B is an example planar array of digital cameras, according to the present invention;

[0027] Fig. 1C is a second example of a linear array of digital cameras, according to the present invention;

[0028] Fig. 2A is an illustration of combined images from multiple cameras and a selection of a "virtual camera" view, according to an embodiment of the present invention;

[0029] Fig. 2B is a block diagram/illustration of some of the possible applications of a process according to an embodiment of the present invention;

[0030] Fig. 3 is an illustration showing example placements and example camera angles of a circular camera array utilized in a teleconferencing application, according to an embodiment of the present invention;

[0031] Fig. 4A is an illustration of a planar camera array, according to an embodiment of the present invention;

[0032] Fig. 4B is an illustration of a planar camera array and example specific views onto a scene, according to an embodiment of the present invention;

[0033] Fig. 4C is a conference/presentation room illustrating views A, B, and C of a camera array, according to an embodiment of the present invention;

[0034] Fig. 5A is an example of an object to be imaged by two separate cameras, according to an embodiment of the present invention;

[0035] Fig. 5B is an example of frames from individual cameras imaging a same object, according to an embodiment of the present invention;

[0036] Fig. 6 is an example of calibration points and patches in raw images and an illustration of a flat panorama constructed from the raw images, according to an embodiment of the present invention;

[0037] Fig. 7 is an example of a warping of a quadrilateral patch ABCD into a square A′B′C′D′, according to an embodiment of the present invention;

[0038] Fig. 8 is an example of images input from 4 sources showing patches, according to an embodiment of the present invention;

[0039] Fig. 9A illustrates an example of an original scene of objects as viewed by two cameras of a camera array, according to an embodiment of the present invention;

[0040] Fig. 9B illustrates an example of the disparity between objects of two combined images when one image is shifted a distance L, according to an embodiment of the present invention;

[0041] Fig. 9C illustrates another example of the disparity between objects of two combined images when one image is shifted a distance L, according to an embodiment of the present invention;

[0042] Fig. 9D illustrates a third example of the disparity between objects of two combined images when one image is shifted a distance L, according to an embodiment of the present invention;

[0043] Fig. 10 is an example of a graph generated using disparity estimation techniques, according to an embodiment of the present invention;

[0044] Fig. 11A illustrates an example of the disparity between objects of two combined images where each image is stretched a distance L, according to an embodiment of the present invention;

[0045] Fig. 11B illustrates a second example of the disparity between objects of two combined images where each image is stretched a distance L, according to an embodiment of the present invention;

[0046] Fig. 11C illustrates a third example of the disparity between objects of two combined images where each image is stretched a distance L, according to an embodiment of the present invention;

[0047] Fig. 12 illustrates an example of stretching an image in two locations, according to an embodiment of the present invention;

[0048] Fig. 13 is an illustration of a cross-fading technique according to an embodiment of the present invention;

[0049] Fig. 14 is an illustration of multiple camera images and quadrilateral regions to be warped from each image, according to an embodiment of the present invention;

[0050] Fig. 15 is a composite panoramic image constructed from the patches (quadrilateral regions) illustrated in Fig. 14 according to an embodiment of the present invention;

[0051] Fig. 16 is a flow diagram illustrating an image composition procedure for each frame in a video sequence, according to an embodiment of the present invention;

[0052] Fig. 17 is a diagram illustrating a radial microphone array, according to an embodiment of the present invention;

[0053] Fig. 18 is a diagram of an automatic camera calibration system using a light source to find registration points, according to an embodiment of the present invention; and

[0054] Fig. 19 is a block diagram of an embodiment of the present invention having a selection device for selecting a view from a combined image.

DETAILED DESCRIPTION

[0055] A "frame" as described herein is a single capture (image) of objects in a field of view of a camera. Motion video consists of a time sequence of frames typically taken at a video rate of ten to thirty frames per second.

[0056] "Parallax" as described herein is the apparent change in the location of an object, caused by a change in observation position that provides a new line of sight. Parallax occurs as a result of two cameras having different COP's. When images from cameras having different COP's are combined without alteration,

disparity between the objects of the combined images results and objects appear "ghosted" or doubled in the composite image.

[0057] A "common field of view" as described herein is an image or a portion of an image of a scene which is viewed by more than one camera.

[0058] Referring now to the drawings, wherein like reference numerals designate identical or corresponding parts, and more particularly to Fig. 1A thereof, there is illustrated an example of a circular array of video cameras, according to an embodiment of the present invention. In Fig. 1A, multiple video cameras 10 are mounted on a rigid substrate 20 such that each camera's field of view overlaps that of its neighbor. The resulting images are aligned using digital warping and objects are matched using disparity reduction techniques to form a large composite image. The result is a seamless high-resolution video image spanning the field of view of all cameras. If the cameras are mounted in fixed positions relative to each other, image registration between the different cameras is also fixed, and the same composition function can be used for each frame. Thus the interpolation parameters need only be calculated once, and the actual image composition and disparity reduction can be done quickly and efficiently, even at video rates.

[0059] Other configurations of cameras are also possible. For example, Fig. 1B illustrates a planar array of cameras 12 mounted on a rigid plane substrate 30, and Fig. 1C illustrates a linear array of cameras 14, each mounted on a rigid substrate 40, according to an embodiment of the invention. The cameras may be aligned in various directions. For example, in Fig. 1B, the cameras are generally aligned in a same direction (with overlapping views as described above), while a

more diverse alignment is seen in Fig. 1C (still generally having overlapping and abutting areas, for example).

[0060] For many applications, such as video conferencing, it is neither desirable nor possible to transmit a full-motion super-resolution image. According to an embodiment of the present invention, a method for extracting a normal-resolution image using measures of motion, audio source location, or face location within the image is provided. However, since multiple cameras and plural images of a same scene are envisioned, super-resolution processes (getting higher resolution, e.g., more pixels, from multiple images of a same scene) may be applied. When the multiple images are registered, any normal-sized sub-image can then be selected by excerpting a frame from the larger composite, as shown in Fig. 2A.

[0061] Fig. 2A illustrates views from 3 cameras, camera 1 view 210, camera 2 view 220, and camera 3 view 230, according to an embodiment of the present invention. A conference participant 200, and the participants frame of reference (an area of interest 240) is at or near abutting images from each of camera 1 view 210 and camera 2 view 220. The abutting areas are combined into a single panoramic image and the area of interest 240 (including participant 200) can be selected therefrom. Camera 3 view 230 may also be combined into the single panoramic image, or, alternatively, because it shows none of the area of interest 240, it may be discarded. The selected combined image 250 (representing the area of interest 240) is the result.

[0062] Because embodiments of the present invention allow any desired sub-image to be selected, changing the selected region is equivalent to steering a "virtual camera." The camera array can be constructed with long-focal-length cameras, such that each array element is "zoomed in" on a small region. Combining many of these zoomed-in images and reducing the resolution is equivalent to zooming out. Thus all the parameters of conventional cameras such as pan (selecting a different area of the panoramic image), zoom (combining or discarding images and changing resolution), and tilt (selecting images from cameras at specific angles, i.e., elevated cameras, for example) can be effectively duplicated with an array of fixed cameras.

[0063] Unlike mechanically-steered cameras which have a finite slew rate limited by motor speed and inertia, a virtual camera can be instantaneously panned anywhere in the camera array's field of view. Multiple virtual cameras can be simultaneously extracted from the same array, allowing multiple users to view different areas at the same time. This might be particularly useful in telepresence applications, for example a sports event, where each user could control a "personal virtual camera."

[0064] For example, one embodiment of the present invention is illustrated in Fig. 2B. A camera array 260 is trained on, for example, a sporting event 262. Video streams from the cameras are then packaged, compressed, and prepared for broadcast at a broadcast station 264, and broadcast via cable, Internet, airwaves (broadcast tower 266, for example), or other broadcasting media/modes. At a receiving end, a receiving device (antennas 280, for example), receive the

broadcast signal, and a user station 282 combines the images received in the broadcast to a single panoramic view. A user (not shown) selects a desired view via control device 284, resulting in a display 286, for example.

[0065] To save broadcast bandwidth, signals from control device 284 may be broadcast back to a source (broadcast station 264 in this example) to identify only the specific video streams needed to produce the display 286. In another alternative the entire display 286 may be composed at the source and only the display 286 is broadcast to the user. A specific configuration would be selected based on the availability of broadcast bandwidth and processing power available at each of broadcast station 264 and user station 282. As will be appreciated by those skilled in the art, any number of combinations or modifications may be implemented, consistent with the invention as described herein.

[0066] Camera arrays can have any arrangement, such as radial (Figs. 1A and 3, for example), linear, planar (Figs. 1B, 1C, and 4, for example), covering the walls of a room, or even attached to the edge of a table or other object. A location and angle of each of the cameras in the camera array is fixed with respect to each other camera. Therefore, the entire camera array may be moved without any re-calibration (re-registration) or recalculation of the matrix equations for transforming the individual images into a single panoramic scene.

[0067] Although normally fixed on a platform or other substrate, it is consistent with an embodiment of this invention to have a camera array with cameras movable with respect to each other and employ a registration process that would re-register each camera in the camera array after movement. In addition,

facilities for additional or auxiliary cameras may be attached or included in the array, and again, a registration process so that any cameras added to the array can benefit from the fast warping/transformation and parallax reduction techniques described herein.

[0068] Referring now to Fig. 3, a radial array of cameras is suitable for video conferencing, having a 360-degree field of view, the radial array is centered on a fixed point, such as the center of a conference table (labeled "B" in Fig. 3). A similar array is useful for immersive telepresence or other applications (discussed below). An array situated at "A," in Fig. 3, gives the impression of being seated at the table.

[0069] A similar array could be mounted on a mobile platform for other telepresence applications. The ability to instantaneously pan without distracting camera motion is an improvement over current telepresence systems. For example, another application, police robotics, require a mechanized camera to proceed into a hostile environment for reconnaissance purposes. Mobile and equipped with a video camera, remote panning means, and microphones, the devices proceed into a hostile environment. When a threat is detected, the camera pans toward the threat to inspect it, however, the device is limited by pan and zoom time parameters. Alternatively, a radial array according to an embodiment of the present invention, mounted on the same robotic device, would be able to instantly pan to evaluate the threat, have multiple redundancy of cameras, and be able to self locate the threat via microphone or motion detection as described hereinbelow.

[0070] Camera arrays, according to an embodiment of the present invention, can be constructed in any desired configuration, and put anywhere convenient. A two-dimensional camera array may be constructed to facilitate electronic zooming as well as panning. For example, a first dimension of an array might cover/or be directed towards a full view of a conference room or scene, while a second array would have cameras directed toward specific points of interest within the scene (see Fig. 1C, for example). Example placement of camera arrays would include an edge of a table, on the ceiling at an angle, or even covering an entire wall or walls of a room. Automatic switching or electronic panning could then ensure an appropriate image of any event in the room, and no part of any scene would ever be out of camera view.

[0071] Fig. 4A shows another application having cameras 400 arrayed in a planar configuration, according to an embodiment of the present invention. Each of cameras 400 is directed to a specific one of abutting regions 410, and a user may select any view covered by any combination of the cameras, allowing up/down panning as well as a good zoom range.

[0072] Fig. 4B illustrates an alternate configuration having cameras 420 (A, B, and C, respectively) directed to specific areas, either abutting or overlapping other camera views, according to an embodiment of the present invention. In this case, specific camera views are directed toward areas of particular interest within a scene (resulting in views A, B, and C, as shown in Fig. 4C, for example). The configuration of Figs. 4B/4C is appropriate for a lecture or conference room application, where the appropriate camera operations is to zoom and pan to follow

the speaker, zoom in on the display screen, and zoom out to cover audience questions or reactions.

[0073] The prior art has failed to adequately resolve the challenging technical constraints of a camera array. Several major problems must be addressed: combining multiple video streams into a seamless composite, calibrating the array cameras, and handling the extreme data rate of the composite high-resolution video.

09650366-121004

Image Warping

[0074] In an embodiment of the present invention, stitching adjacent frames together is accomplished using a method or combination of methods that combine separate images into a panoramic, or combined image. In one embodiment, a spatial transformation (warping) of quadrilateral regions is used, which merges at least two images into one larger image, without loss of generality to multiple images. First, a number of image registration points are determined; that is, fixed points that are imaged at known locations in each sub-image. This can be done either manually or automatically. In either case the process involves pointing the array at a known, structured scene and finding corresponding points. For example, Fig. 5A illustrates views of two cameras trained on a scene that includes a rectangular box 500. The rectangular box 500 is in an area of overlap between a View 1 and View 2 (each View 1 and View 2 corresponding to an approximate field of view captured from one of the two cameras). Therefore, each of the points

of rectangular box 500 would constitute image registration points (points in common to views of each of two or more cameras of the camera array).

[0075] Fig. 5B illustrates the rectangular box 500 as captured in an actual frame of the two cameras corresponding to View 1 and View 2 of Fig. 5A, as abutting frames, frame 1 and frame 2. The box is a quadrilateral area EFGH 510 in frame 1, and E′F′G′H′ 520 in frame 2, and, because of the slightly different camera angles, the quadrilateral areas are not consistent in angular construction and are captured in different locations with respect to other objects in the image. Therefore, the present invention matches the abutting frames by warping each of the quadrilateral regions into a common coordinate system. Note that the sides of quadrilateral area EFGH are shown as straight, but may actually be subject to some barrel or pincushion distortion, which can also be approximately corrected via the fast warping equations (discussed below). Barrel/pincushion distortion can be corrected using radial (rather than piecewise linear) transforms. Piecewise linear transforms can fix an approximation of the curve.

[0076] Alternatively, only one of the images need be warped to match a coordinate system of the other image. For example, warping of quadrilateral area EFGH may be performed via a perspective transformation. Thus quadrilateral EFGH in Frame 1 can be transformed to E′F′G′H′ in the coordinate system of Frame 2.

[0077] In another embodiment, bilinear warping (transformation) of piecewise-contiguous quadrilateral regions is used. Referring now to Fig. 6, an example of how multiple images are merged into one larger image, without loss of

generality to the multiple images (also referred to as sub-images) according to an embodiment of the present invention is described. First, a number of image registration points are determined; that is, fixed points that are imaged at known locations in each sub-image. In this embodiment, the fixed points at known locations consist of a cylinder 600 coaxial with a camera array axis 610, and having a radius 620 of about the working range of the camera (in an embodiment, about a meter). A wide angle camera is used having a small f stop, thereby providing a large depth of field where everything beyond approximately 1 meter is in focus.

[0078] Square patches on the cylinder (640, for example) are imaged as quadrilateral regions by one or more cameras. The imaged quadrilateral regions are illustrated, for example, as quadrilateral region ABCD as seen in Fig. 7, a patch, and may also be referred to as a source polygon. Each quadrilateral region (source polygon, or patch) is warped back to a square, and the final panoramic image 650 (also referred to as a destination image) is composited by abutting each square in a grid (also referred to as placing the squares, or warped quadrilateral regions, into a common coordinate system). Different camera images are merged by either abutting adjacent squares in a grid or combining overlapping squares (described below) resulting in panoramic image 650.

[0079] Bilinear transformations may be used to warp the quadrilateral regions (bilinear warping). Equation 1 below transforms the homogeneous coordinate system u, v to the warped (square) coordinate system x, y.

$$[x \;\; y] = [u \;\; v \;\; uv \;\; 1] \begin{vmatrix} a0 & b0 \\ a1 & b1 \\ a2 & b2 \\ a3 & b3 \end{vmatrix} \qquad (1)$$

[0080] Equation 1 is a transformation matrix having 8 unknown coefficients that are determined by solving the simultaneous equations given by the reference points (ABCD, in a coordinate system u, v, and A′B′C′D′, in a coordinate system x, y). The four points in each system have 8 scalar values to solve for the 8 unknown parameters. If more correspondence points (correspondence points referring to the points encompassing the quadrilateral region (ABCD in this example) are present, an overdetermined set of equations results, which can be solved using least-squares (pseudoinverse) methods for more robust estimates.

[0081] The above processes may be repeated for every patch (each patch captured by one or more cameras; at least two sets of patches along borders of images one from each image to be combined; each path determined along a border of combined images), and using equation (1) (or an equivalent equation performing the same function for the areas selected), a set of warping coefficients (eight coefficients in this example) are computed for every patch. These, as well as the location of the square destination region in the composite image, are referred to as warping coefficients or a calibration set.

[0082] To calculate a pixel value in the warped coordinate system x, y, the above equations are inverted by solving for u, v in terms of x, y. This allows for what is termed "inverse mapping." For every pixel in the warped coordinate system, the corresponding pixel in the unwarped system is found and its value is copied.

[0083] The coefficients (of equation 1, for example) are stored in a table and utilized to warp images "on the fly." Because the cameras are fixed, and registered (or calibrated, see later section), the same equations are utilized over and over for the same patches (i.e., no need to find new correspondence or registration points, or recalculate coefficients).

[0084] Because warping is a continuous function rather than discrete, the reverse mapping will generally yield non-integral unwarped coordinates. For this reason, the pixel value is interpolated from the neighbors using bilinear interpolation. This uses a linear combination of the four closest integral points to produce the interpolated value.

[0085] Other embodiments include different types of spatial transformations to warp patches from captured images (u,v coordinate system) to a composite grid (x,y coordinate system). Any spatial transformation altering the captured images to fit into a composite grid would be consistent with the present invention. For example, affine, or perspective transformations may be utilized.

[0086] According to one embodiment, registration is performed manually by inspecting each camera's image. This is not an excessive burden as it need only be done once, and can be automated.

[0087] Fig. 8 provides an example of calibration points and patches in raw images (images taken from a camera, without having any warping or other processing applied), according to an embodiment of the invention. Image 810, 820, 830 and 840 are abutting images taken from consecutively aligned cameras on a scene. Each image has a grid 815, 825, 835, and 845 having source polygons

that will be warped into a common coordinate system resulting in a contiguous image of the scene. In addition, hatched columns 851 and 876 (seen in both camera images) are combined, while column 877 and column 852 may be either utilized as is, discarded, or combined.

[0088] An embodiment of the present invention also includes correction for lens distortions in the fast warping equations. For example the camera lenses utilized in Fig. 8 shows a substantial amount of barrel distortion. Note the right edge of partition 850 shown on the right side of image 810 (bulging to the right), and the right edge of partition 850 shown in the left side of image 820 (bulging to the left). In one embodiment, correction of lens non-linearities or distortions is built into the fast warping equations (using a radial transformation, for example). In another embodiment, such abnormalities are corrected by increasing the number of patches (registration points), and the resulting warped images making a better approximation of the actual scene being imaged (a piecewise approximation of a continuous distortion).

FXPL-01032US0

Parallax Compensation

[0089] Because it is impractical to make a camera array with each camera having a common COP, camera arrays will have a small but significant baseline separation. This is a problem when combining images of objects at different distances from the baseline, as a single warping function will only work perfectly for one particular distance. Images of objects not at that distance may be warped

into different places – resulting in image disparity – and may appear doubled ("ghosted") or truncated when the images are merged.

[0090] According to an embodiment of the present invention, image disparity can be reduced by adjusting the images to determine a minimum disparity between the images being merged. In this embodiment, the camera array is calibrated such that objects at a particular distance, or images of smooth backgrounds, can be combined with no visible disparity. A minimum disparity can be found by determining how much to shift one image to match the other. This type of camera calibration greatly simplifies the stereo matching problem, turning it into essentially a one-dimensional search. Because images are warped into corresponding squares, all that is necessary is to find a particular shift that will match them.

[0091] Figure 9A illustrates a scene 900 containing objects 901, 903, and background 905. When scene 900 is viewed using a multiple-camera viewing system having two cameras with different COP, a left camera and a right camera as described above, the warped images 900L and 900R of scene 900 appear different. Image 900L illustrates the scene 900 as viewed by a left camera after the image has been warped into a common coordinate system. Image 900L includes images 901L, 903L and background 905L of objects 901, 903 and 905 as viewed by a left camera array. Image 900R illustrates the view of scene 900 as it is viewed from a right camera of a camera array after the image has been warped. Image 900R contains images 901R, 903R and background 905R of objects 901, 903 and

905. Directly overlaying the images 900L, 900R results in a combined image similar to that of image 910 as illustrated in Figure 9A.

[0092] As can be seen in image 910, images 901L and 901R are not lined up with one another due to the way scene 900 is viewed by cameras with different COP. Similarly, images 903R and 903L are likewise not lined up exactly with one another. However as can be seen, the combined background images 905R and 905L, which correspond to the original background 905, do not result in any discernable disparity because the background is uniform throughout the scene 900.

[0093] In an embodiment, disparity may be reduced by shifting one or both of the images 900L, 900R so that disparity between images 901L, 901R, 903L, 903R is reduced. To determine how far to shift the image or images, a disparity estimation technique, or a combination of disparity estimation techniques, may be performed to determine disparity differences between the images at different amounts of shift.

[0094] Disparity estimation techniques may be used to determine the disparity between objects imaged in two images, such as images 901L and 901R of images 900L and 900R. According to an embodiment of the present invention, any number of disparity estimation techniques, or combination of techniques, may be used to generate a numerical disparity value.

[0095] One such disparity estimation technique is the pixel-wise difference magnitude technique ("PDM"). PDM calculates the sum of the difference of all pixel values of two corresponding images, such as images 900L and 900R. The

pixel difference is typically squared, or the absolute value is taken, to result in a positive magnitude before summation.

[0096] Another disparity estimation technique is known as normalized pixel difference ("NPD"). NPD determines disparity by normalizing or scaling the pixels from both source images to correct for camera intensity or color differences. Then the difference magnitude may be calculated using another disparity technique, such as PDM. **[0097]** Still another technique for determining disparity between images utilizes an edge detector or other gradient or linear operation before the difference magnitude is computed, as described above.

[0098] Another disparity estimation technique is known as color or intensity histogram differences ("CIHD"). Using this technique, the count of pixels having intensities within specified bins is used to characterize each image region. These histogram counts may be subtracted and a difference magnitude computed.

[0099] Another technique is frequency-domain comparisons ("FDC"). Using this technique, a spatial transform such as a two-dimensional Fourier transform or wavelet transform may be used to parameterize image regions into transform coefficients. The magnitude difference can be computed between the transform coefficients, which can be selected or weighted to favor or discard particular spatial frequencies.

[0100] Statistical comparisons may also be used to compare the image regions, for example the distribution of pixel intensities can be modeled with a multivariate Gaussian distribution for an image region. Gaussian distributions can

be compared using one of several distance measures, such as the Karhunen-Loeve measure.

[0101] Still another technique is known as feature-based comparisons ("FBC"). According to this technique, image regions may be compared by extracting any number of features, such as connected regions or lines having similar color or texture. The features may be compared by size, color, shape, spatial orientation, location, or any other thing that results in a measure of image disparity.

[0102] Using any of the above disparity estimation techniques or a combination thereof, overlapping images 900L and 900R may be compared a multitude of times at different overlapping ranges (image adjustment), thereby generating a graph such as the one illustrated in Figure 10 showing the difference in disparity associated with the amount of image adjustment L. L may be any numerical value for adjusting, or shifting images, such as: pixel, millimeter, centimeter, inch, etc. Additionally, L does not need to be an integer; it can be any real number.

[0103] Figure 9B illustrates the disparity between images 900L and 900R, where L=0. The disparity diagram 909 illustrates the pixels having the same or similar values as dark gray 907. The pixel values having a different value due to the disparity are illustrated as light gray areas 901′ and 903′. Using one of the disparity estimation techniques, a value is assigned to represent the pixel difference at L=0.

[0104] Figure 9C illustrates images 900L, 900R combined where image 900L has been adjusted, or shifted, to the right with reference to image 900R an L distance equal to 1. As can be seen, the resulting pixel difference regions 901″ and 903″ illustrated in light gray are less than when compared to a disparity shown in Figure 9B.

[0105] Likewise, Figure 9D illustrates the disparity 901‴ and 903‴ in images 900L and 900R when the image 900L has been shifted to the right a length equal to 3.

[0106] Disparity estimation techniques can be run a multitude of times for different values of L, thereby generating a graph, such as the graph illustrated in Figure 10, according to an embodiment of the present invention. Within the graph there will be a minimum disparity value for a specific adjustment ΔL, which can be selected to use when combining overlapping images to generate a composite panoramic image.

[0107] In an embodiment, ΔL may be computed by shifting images with relationship to one another, shifting both images, or shifting one image with relation to the other a multitude of times and computing disparity using one of the above-described disparity estimation techniques for each value of L.

[0108] In an embodiment, disparity estimation/correction may be calculated for every frame, or adjusted based on need. For example, calculation and selection of ΔL may only be computed every other frame, every fifth frame, every other second, or any other selected re-computation frequency. For frames where ΔL is

not computed, the then-existing value of ΔL may be used to adjust and combine the images.

[0109] Similarly, in an embodiment, instead of calculating disparity for every possible value of L, L is only computed a selected number of times and the minimum L from the computations is used as ΔL.

[0110] In still another embodiment, instead of computing the disparity for each entire image, only a portion of each combined image is sampled and computed. For example, to speed computation, every third pixel row of the combined images is compared for calculating disparity and finding ΔL, thereby reducing the overall computation time.

[0111] Additionally, images may be divided into multiple patches and disparity may be computed for different values of L for each overlapping patch and each overlapping patch adjusted according to its own selected value of ΔL.

[0112] Any technique, or combination thereof, may be used for determining ΔL.

[0113] In an alternative embodiment, disparity between combined images may be corrected by "stretching" one or both images to reduce the difference between overlapping images of objects. Similar to shifting images, determining how far to stretch the image or images may be performed using one, or a combination of the above-described disparity estimation techniques.

[0114] Referring to Figure 11A, using any of the above disparity estimation techniques or a combination thereof, multiple images, such as images 1100L and 1100R may be compared a multitude of times at different stretch lengths L, thereby

generating a graph similar to the one illustrated in Figure 10 showing the disparity associated with the amount of stretching L. L may be any numerical value such as: pixel, millimeter, centimeter, etc. Additionally, L does not need to be an integer, it can be any real number.

[0115] Figure 11B illustrates the difference resulting from combining images 1100L and 1100R, without stretching (L=0). The difference diagram 1109 illustrates pixels having the same or similar values as dark gray 1107. The pixel values having a difference value due to a parallax problem are illustrated as light gray areas 1101′ and 1103′.

[0116] Figure 11C illustrates images 1100L and 1100R overlapped where image 1100L has been stretched to the right with reference to image 1100R an L distance equal to 1, and image 1100R has been stretched to the left with reference to image 1100L an L distance of 1.

[0117] In an embodiment, stretching an image is performed by keeping a portion of an image stationary, for example, one side of the image, while the remainder of the image is stretched. For example, referring again to Figure 11A, in an embodiment, image 1100L is stretched to the right while the left side of the image remains stationary. Therefore the change in pixel values would be greater for the pixels on the right portion of the image in comparison to the pixels on the left portion of the image.

[0118] In an alternative embodiment, a point or particular portion of an image may be kept stationary and the remainder of the image stretched from there

outward. This embodiment may be useful when one portion of the image is being focused on and/or is not overlapping any other camera view.

[0119] Additionally, when combining images using stretching techniques, one image may be stretched while the others remain unaltered, or all images may be stretched.

[0120] Stretching images, as described above, provides the ability to correct multiple occurrences of disparity in a single image. For example, referring to Figure 12, scene 1200 containing objects 1211 and 1213 is viewed by three cameras of a multiple-camera viewing system, each camera having camera views 1201, 1203 and 1205, respectively. As can be seen the camera view 1201 has two overlaps 1207 and 1209, where more than one camera shares a common field of view. By selecting the center line of image 1201 as the stationary point, either side of image 1201 may be stretched, as described above, with respect to the center line to reduce disparity created in the common field of view 1209 of camera views 1201 and 1203, and the common field of view 1207 of camera views 1201 and 1205. In this embodiment, the different common fields of view 1209, 1207 of image 1201 may be stretched different amounts to compensate for parallax between the three cameras. Additionally, the common field of view 1209 from camera 1203 and the common field of view 1207 from camera 1205 may also be adjusted to help correct the disparity.

[0121] In an embodiment, disparity estimation/correction may be calculated for every frame, or adjusted based on need. For example, calculation and selection of ΔL may only be computed every other frame, every fifth frame, every other

second, or any other selected re-computation frequency. For frames where ΔL is not computed, the then-existing value of ΔL may be used to adjust and combine overlapping images.

[0122] Similarly, in an embodiment, instead of calculating disparity for every possible value of L, L is only computed for a selected number of times and the minimum L from the computations is used as ΔL.

[0123] In still another embodiment, instead of computing the disparity for each entire image, only a portion of each combined image is sampled and computed. For example, to speed computation, every third pixel row of the combined images is compared for calculating disparity, thereby reducing the overall computation time. Alternatively, only the portion of each image which shares a common field of view is calculated.

[0124] Additionally, images may be divided into multiple patches and disparity may be computed for different values of L for each overlapping patch and each overlapping patch adjusted according to its own selected value of ΔL based on its disparity estimate.

[0125] Any technique, or combination thereof for calculating ΔL may be used.

[0126] In addition to correcting disparity between overlapping portions of images, disparity estimates have a number of additional applications. For example, disparity estimates may be used to estimate the location and number of objects or humans in the array's field of view. This is done by looking for local peaks in the smoothed disparity map (a smoothing of the grid (Fig. 10) of disparity estimates

for each patch), using template matching or a similar technique. Once the positions are estimated, the virtual cameras can be "snapped to" the object locations ensuring they are always in the center of view.

[0127] If it is desired to find the range of moving objects such as humans, the above techniques can be used on the frame-by-frame pixel difference for more robust disparity estimates.

[0128] Disparity depends directly on the distance of the object in a patch from the cameras. This will not have high resolution due to the small camera baseline, and will often be noisy, but still is able to detect, for example, the position of humans sitting around a conference table. Patches may be small, on the order of 10-50 pixels, and can be overlapped for greater spatial detail, as there will be only one disparity estimate per patch. The result is a grid of disparity estimates and their associated confidence scores.

[0129] Another application is to ensure that camera seams do not intersect objects. Because the individual camera views may overlap by several patches, the system can be configured to blend only those patches that do not have significant disparity estimates, that is, contain only background images instead of objects.

[0130] Yet another application is to segment the panoramic images into foreground and background images. Once again, this task is simplified by the fact that the camera array is fixed with respect to the background. Any image motion is thereby due to objects and objects alone. Patches with no significant motion or disparity are background images. As long as foreground objects move enough to reveal the background, the background can be robustly extracted. The difference

between any given image and the background image will be solely due to foreground objects. These images can then be extracted and used for other applications; for example, there is no need to retransmit the unchanging background. Significant bandwidth savings can be gained by only transmitting the changing object images (as recognized in the MPEG-4 video standard).

Cross-Fading

[0131] Once the source polygons have been warped to a common coordinate system and parallax corrected for overlapping images, a "cross-fade" can be utilized to combine them, as illustrated in Fig. 13.

[0132] In this process, two or more images are obtained from different cameras. For example, cameras 1300 and 1320 of Fig. 13 each have imaged a same square 1310. Each image is faded by reducing the pixel intensities across the patch (imaged square) in a linear fashion. For example, the patch from the left camera 1300 is faded so that pixel intensities fade to zero at the rightmost edge of the patch, while the leftmost are unchanged (see expanded patch 1330). Conversely, the right camera patch is faded such that the leftmost pixels go to zero (see expanded patch 1340). When the pixel intensities of right and left patches are added, the result is a patch image that smoothly blends the two images. This technique further reduces artifacts due to camera separation and image intensity differences.

[0133] Patches for fading may be of any size or shape within the confines of a camera view. Therefore any geometric shape or outline of an object or other selected area within two or more camera views may be selected for fading.

[0134] The number and size of patches can be adjusted as necessary to give a good mapping and reduce the appearance of "seams" in the composite image. "Seams" can be put at arbitrary locations by changing the interpolation function, for example, by using a piecewise linear interpolation function. This is especially useful when face locations are known, because the image combination can be biased such that seams do not cut across faces.

[0135] Alternatively, individual corresponding pixels in overlapping regions, after warping and matching those images, may be summed and averaged in some manner and then faded or blurred at edges or throughout the overlapping regions. As will be appreciated by those skilled in the art, many procedures, including fading, blurring, or averaging may be implemented to smooth transitions between the warped abutting and/or combined images.

[0136] Figs. 14 and 15 show how multiple images can be integrated into a high-resolution composite. Fig. 14 illustrates images taken from cameras 1400 (CH1), 1410 (CH2), 1420 (CH3), and 1430 (CH4). Each of the images includes a set of quadrilateral grids to be warped into a common coordinate system, according to an embodiment of the present invention.

[0137] Fig. 15 shows a composite 1500 of the images of Fig. 14 after warping, disparity correction, cross-fading and combined into a common coordinate system, according to an embodiment of the present invention. Fig. 15

includes a grid 1510 corresponding to the common coordinate system, and is provided for reference. Note how the images are not only combined, but distortions and parallax are corrected. For example, the curved edges of the leftmost wall (along tiles A1..A5) are straight in the composite and there is no ghosting effect from combined images.

[0138] Also note the effects of cross-fading to produce a seamless image. Quadrilateral regions 1402 and 1412, and 1404 and 1414 (Fig. 14) of the camera images are combined to produce grid squares E2 and E3 of the composite. Quadrilateral regions 1402 and 1404 of CH1 are darker than corresponding regions 1442 and 1414 of CH2, as is common in similar views from different cameras. However, when combined the rightmost portions of grid squares E2 and E3 are light (as in quadrilateral regions 1412 and 1414), while the leftmost regions of grid squares (E2 and E3) are dark (as in quadrilateral regions 1402 and 1404), and no seams are present.

[0139] As one who is skilled in the art would appreciate, Figure 16 illustrates logic steps for performing specific functions. In alternative embodiments, more or fewer logic steps may be used. In an embodiment of the present invention, a logic step may represent a software program, a software object, a software function, a software subroutine, a software method, a software instance, a code fragment, a hardware operation or user operation, singly or in combination.

[0140] Fig. 16 is a flow diagram illustrating the steps for compositing each video frame. In this flow, three cameras (1600-1, 1600-2, and 1600-3) each provide one image to be combined as part of a composite panoramic image. At

steps 1610-1, 1610-2, and 1610-3, each of the images are processed by selecting patches (quadrilateral regions) and warped into a space for fitting into a common coordinate system. At steps 1615-1, 1615-2, and 1615-3, parallax from overlapping quadrilateral regions (patches) is corrected using image disparity estimation techniques, and at step 1620, the patches are cross-faded to eliminate edges and seams, and placed in tiles (the common coordinate system, for example). Upon completion, the composite panoramic image is available for display, selection, storage or other processes (step 1630). Note that this procedure, including all data transfer and warping, using currently available processing speeds, can be performed at video rates of 10-30 frames per second.

Automatic Control of Virtual Cameras

[0141] Mechanically-steered cameras are constrained by the limitations of the mechanical systems that orient them. A particular advantage of virtual cameras is that they can be panned/zoomed virtually instantaneously, with none of the speed limitations due to moving a physical camera and/or lens. In addition, moving cameras can be distracting, especially when directly in the subject's field of view, like the conference-table camera shown in Fig. 3.

[0142] In this system, we can select one or more normal-resolution "virtual camera" images from the panoramic image. Mechanical cameras are constrained by the fact that they can be pointed in only one direction. A camera array suffers no such limitation; an unlimited number of images at different pans and zooms can be extracted from the panoramic image. Information from the entire composite

image can be used to automatically select the best sub-images using motion analysis, audio source location, and/or face tracking. To reduce the computation load, parts of the panoramic image not used to compose the virtual images could be analyzed at a slower frame rate, resolution, or in greyscale.

[0143] A useful application of camera arrays is as a wide-field motion sensor. In this case, a camera array is fixed at a known location in a room. Areas of the room will correspond to fixed locations in the image plane of one or more cameras. Thus using a lookup table or similar method, detecting motion in a particular region of a video image can be used to find the corresponding spatial location of the motion. This is enough information to point another camera in the appropriate direction, for example. Multiple cameras or arrays can be used to eliminate range ambiguity by placing their field of view at right angles, for example, at different room corners.

[0144] Another useful system consists of conventional steerable cameras and a camera-array motion sensor. Motion in a particular location would set appropriate camera pan/zoom parameters such that a subject is captured. For example, in Fig. 4C, motion above the podium would signal the appropriate camera to move to a location preset to capture a podium speaker. This mode of operation is computationally cheap, and less expensive black-and-white ("B/W") cameras could be used in the camera array, as the resultant image need not be shown. This could have significant savings in processing as well, as three B/W cameras could be multiplexed on one RGB video signal.

Camera Control Using Video Analysis

[0145] In one embodiment, a motion analysis serves to control a virtual camera; that is, to select the portion of the panoramic image that contains the moving object. Motion is determined by computing the frame-to-frame pixel differences of the panoramic image. This is thresholded at a moderate value to remove noise, and the center of gravity (first spatial moment) of the resulting motion image is used to update the center of the virtual image. The new virtual image location is computed as the weighted average of the old location and the motion center of gravity. The weight can be adjusted to change the "inertia," that is, the speed at which the virtual camera changes location. Giving the previous location a large weight smooths jitter from the motion estimate, but slows the overall panning speed. A small weight means the virtual camera responds quickly to changes in the motion location, but may jitter randomly due to small-scale object motion.

[0146] Tracking can be further improved by adding a hysteresis value such that the virtual camera is changed only when the new location estimate differs from the previous one by more than a certain amount. The motion centroid is averaged across both a short and a long time span. If the short-time average exceeds the long-time average by a preset amount, the camera view is changed to that location. This accounts for "false alarm" events from both stable sources of image motion (the second hand of a clock or fluorescent light flicker) as well as short-term motion events (such as a sneeze or dropped pencil). This smooths jitter, but

constant object motion results in a series of jumps in the virtual camera position, as the hysteresis threshold is exceeded.

[0147] Other enhancements to the motion detection algorithm include spatial and temporal filtering, for example, emphasizing hand gestures at the expense of nodding or shifting. In operation, the virtual camera is initially zoomed out or put in a neutral mode, which typically includes everything in the camera's view. If a radial array is used, as in Fig. 3, the composite image will have a narrow aspect ratio, that is, will be wider than it is high. In this case, the neutral view can be "letterboxed" into a normal-aspect frame by reducing it in size and padding it with black or another color.

[0148] Alternatively, the neutral position could be a "Brady Bunch" view where the large image is broken into units to tile the normal-aspect frame. The output of a face tracker ensures that all participants are in view, and that the image breaks do not happen across a participant's face.

[0149] If motion is detected from more than one region, several heuristics can be used. The simplest is to just choose the region with the largest motion signal and proceed as before. Another option might be to zoom back the camera view so that all motion sources are in view. In the case of conflicting or zero motion information, the camera can be changed back to the default neutral view.

[0150] Another useful heuristic is to discourage overlong scenes of the same location, which are visually uninteresting. Once the virtual camera location has been significantly changed, a timer is started. As the timer value increases, the motion change threshold is decreased. This can be done in such a way that the

mean or the statistical distribution of shot lengths match some pre-determined or experimentally determined parameters. Another camera change resets the timer. The net effect is to encourage human-like camera operation. For example, if the camera has been focused on a particular speaker for some time, it is likely that the camera would cut away to capture a listener nodding in agreement, which adds to the realism and interest of the video, and mimics the performance of a human operator.

[0151] All the above techniques can be combined with the object locations estimated from the disparity map.

Audio Control of Virtual Cameras

[0152] Using microphone arrays to determine the location and direction of acoustic sources is known. These typically use complicated and computationally intensive beamforming algorithms. However, a more straightforward approach may be utilized for determining a direction of a speaker at a table, or from which side of a room (presenter or audience) speech is coming from. This information, perhaps combined with video cues, a camera can be automatically steered to capture the speaker or speakers (using the methods of the previous sections).

[0153] While conventional beamforming relies on phase differences to estimate the direction and distance of an acoustic source, an embodiment of the present invention obtains a good estimate using the amplitude of an acoustic signal. The system presented here uses an array of directional microphones aligned around a circle as shown in Fig. 17. Such an array is placed in the center of a meeting

table or conference room. If the microphones are sufficiently directional, then the microphone with the highest average magnitude should indicate the rough direction of an acoustic source.

[0154] Adverse effects, such as acoustic reflections (not least off walls and tables), and the imperfect directionality of real microphones (most cardioid microphones have a substantial response at 180 degrees to their axis) are minimized according to an embodiment of the present invention.

[0155] In one embodiment, the present invention utilizes a pre-filtering of the acoustic signal to frequencies of interest (e.g. the speech region) helps to reject out-of-band noise like ventilation hum or computer fan noise.

[0156] In addition, lateral inhibition is utilized to enhance microphone directionality. In one embodiment, lateral inhibition is done by subtracting a fraction of the average signal magnitude from each neighboring microphone. The time-averaged magnitude from each microphone is denoted |M| as illustrated in Fig. 17. A small fraction $\alpha < 1$ (α 1700 and α 1701 subtracted from Mn, for example) is subtracted from each of the neighbor microphones. This sharpens the spatial resolution from each microphone. The neighbor dependence can be increased beyond nearest neighbors if necessary. If the subtraction is done in the amplitude or energy domain, then problems due to phase cancellation and reinforcement are avoided altogether. The result is to sharpen the directionality of each microphone.

[0157] In one embodiment, the system is normalized for ambient conditions by subtracting the ambient energy incident on each microphone due to constant sources such as ventilation. When the system is running, each microphone will

generate a real-time estimate of the acoustic energy in its "field of view." It might be possible to get higher angular resolution than the number of microphones by interpolation.

[0158] A more robust system estimates the location of an acoustic source by finding peaks or corresponding features in the acoustic signals from each microphone. Because of the finite speed of sound, the acoustic signal will arrive first at the microphone closest to the source. Given the time delay between peaks, the first-arriving peak will correspond to the closest microphone. Given delay estimates to microphones at known locations, geometrical constraints can be used to find the angular direction of the source.

[0159] In a complex environment with many reflections, the statistics of reflections may be learned from training data to characterize the angular location of the source. Combining this with the amplitude cues above will result in an even more robust audio location estimate.

[0160] A system with particular application to teleconferencing consists of one or more desk microphones on flexible cords. In use, microphones are placed in front of each conference participant. Each microphone is equipped with a controllable beacon of visible or invisible IR light. The beacon is set to flash at a rate comparable to ½ the video frame rate. Thus there will be frames where the beacon is illuminated in close temporal proximity to frames where the beacon is dark. Subtracting these frames will leave a bright spot corresponding to the beacon; all other image features will cancel out. From this method the location of the microphones in the panoramic image can be determined. Audio energy

detected at each particular microphone can give a clue to shift the virtual image to that microphone.

Camera Control Using Audio

[0161] Given the angular direction and the magnitude signals from the various microphones, cameras can be controlled using algorithms similar to those described for motion above according to an embodiment of the present invention. Short-term and long-term averages can accommodate fixed noise sources like computer fans. A number of heuristics are used to integrate face and/or motion detection with the audio source location. The audio direction is used as an initial estimate of speaker location to start the face tracking system. Additionally, both face tracking and audio source location can be used in concert. Thus an object must both be recognized as a face and be an audio source before the camera is steered towards it. This is particularly useful for the automatic teleconferencing system that aims to display the image of the person speaking.

Stereo Ranging

[0162] Since the present invention utilizes multiple cameras, stereo ranging of objects in the scene to be imaged may be performed. Conference room walls would be considered at a maximum range, but any object closer to the camera would be considered to be more likely subjects for moving focus of a virtual view of the image. If stereo ranging determines an object is closer than the conference room walls, and it is moving (motion detection via video analysis, for example) it

is more likely to be a subject. If audio detection is also added, the object can be determined with even a higher degree of certainty to be a subject for zooming in. In an embodiment, the present invention utilizes an embodiment utilizing a combination of all analysis functions, audio, video motion detection, and stereo ranging to determine a likely subject for camera zooming.

Data Manipulation and Compression

[0163] Multiple video cameras require techniques to cope with the sheer amount of generated video data. However, it is quite possible to composite multiple video streams in real time on a common CPU, and this should scale with increasing processor speed and parallelization. It is possible to stream each camera to a plurality of analog or digital recording devices, such that all camera views are recorded in real time. The recorded streams can then be composited using the same methods at a later time. Another approach is to store the composited high-resolution video image in a format that can support it.

[0164] Many common video formats such as MPEG support arbitrarily large frame sizes. Recording a full-resolution image has many advantages over prior art: first of all multiple views can still be synthesized from the high-resolution image, which may support varied uses of the source material. For example, in a videotaped lecture, one student might prefer slide images while a hearing-impaired but lip-reading student might prefer the lecturer's image. Recording a full-resolution image also allows better automatic control. Because any real-time camera control algorithm can't look ahead to future events, it is possible to get

better control using a lag of several seconds to a minute. Thus switching to a different audio or motion source could be done instantaneously rather than waiting for the short-term average to reach a threshold.

[0165] Existing standards like MPEG already support frames of arbitrary resolution provided they are rectangular. It is possible to composite images using MPEG macroblocks rather than in the pixel domain for potentially substantial savings in both storage and computation. The multi-stream approach has the advantage that only the streams needed for a particular application need be considered.

[0166] For example, when synthesizing a virtual camera from a circular array ("B" of Fig. 3, for example), a likely application would only require, at most, two streams of video to be considered at any one time. When combined with the fast warping techniques associated with the present invention, such processing is well within the capacity of a desktop PC.

[0167] A reverse embodiment is also envisioned: extracting normal-resolution video from a super-resolution MPEG stream is merely a matter of selecting and decoding the appropriate macroblocks. Given bandwidth constraints, a panoramic video image may be efficiently transmitted by sending only those regions that have changed significantly. This technique is commonly used in low-bandwidth video formats such as H.261. A novel adaptation of this method is to only store or transmit image regions corresponding to moving faces or a significant audio source such as a speaker.

Automatic Camera Registration

[0168] In order to merge overlapping images from different cameras, they must be registered such that lens and imaging distortion can be identified and corrected. This is particularly important with embodiments of the present invention that utilize the matrix coefficients, as they are premised on registered cameras. Generally, it is envisioned that cameras in the arrays will be fixed with respect to one another, and that registration will be performed at time of manufacture.

[0169] The present invention includes registering array cameras that are fixed with respect to each other. Registering cameras involves finding points that correspond in each image. This can be done manually, by observing two views of the same scene and determining which pixels in each image correspond to the same point in the image plane. Because cameras are fixed with respect to each other, this need only be done once and may be performed automatically. Manual registration involves locating registration points manually, say by pointing the camera array at a structured image such as a grid, and locating grid intersection points on corresponding images. Using machine-vision techniques, this could be done automatically.

[0170] In one embodiment, registration is performed using a "structured light" method (e.g. a visible or IR laser spot swept over the camera array's field of view, as shown in Fig. 18). In this example, a semiconductor laser 1800 (other types of light sources may be utilized, a focused infrared beam, for example) is arranged to project a bright spot 1810 of visible red or infrared light on a scene

1820 to be imaged. An image from each camera (camera 1801 and 1802, in this example) is then analyzed to detect the spot location, which then serves as a registration point for all cameras that have it in view.

[0171] Because the spot 1810 is orders of magnitude brighter than any projected image, detection can be performed by thresholding the red channel of the color image (other detection methods are also envisioned, color differences, or analyzing a combination of shape and brightness, for example). The spot 1810 also needs to be moved to find multiple registration points. This could be done using a rotating mirror or other optical apparatus, using multiple lasers (which are inexpensive), or by affixing a laser to a mechanically steered camera as described previously.

[0172] Another version of this system uses bright IR or visible LEDs affixed to a rigid substrate. Lighting the LEDs in succession provides registration points. The substrate can be moved to the approximate imaging error so that parallax is minimized at those points.

[0173] One embodiment of the present invention is illustrated in the block diagram of Fig. 19. Fig. 19 illustrates a video system 1900 having a camera array 1910, a combining device 1930, a view selection device 1960, and an output mechanism 1970. The camera array 1910 is trained on a conference lecture 1920. Images from the camera array 1910 are combined in a combining device 1940 using any of the above processes (warping via a warping mechanism 1980, correcting parallax via a parallax correction mechanism 1985, and cross fading via a fading device 1990, for example). In this embodiment, the combined image is

stored in a memory 1950, and the view selection device 1960 selects a part of the combined image for display on output mechanism 1970. The view selection device may make its selection based on inputs from a user input via input mechanism 1975 (a trackball, mouse or keyboard, for example), or it may automatically select a view based on the above discussed audio inputs, stereoscopic ranging, or video motion analysis.

[0174] The present invention may be conveniently implemented using a conventional general purpose or a specialized digital computer or microprocessor programmed according to the teachings of embodiments of the present invention.

[0175] An embodiment of the present invention includes a computer program product which is a storage medium (media) having instructions stored thereon/in which can be used to program a computer to perform any of the processes of an embodiment of the present invention. The storage medium may include, but is not limited to, any type of disk including floppy disks, optical discs, DVD, CD-ROMs, microdrive, and magneto-optical disks, ROMs, RAMs, EPROMs, EEPROMs, DRAMs, VRAMs, flash memory devices, magnetic or optical cards, nanosystems (including molecular memory ICs), or any type of media or device suitable for storing instructions and/or data.

[0176] Stored on any one of the computer readable medium (media), an embodiment of the present invention may include software for controlling both the hardware of the general purpose/specialized computer or microprocessor, and for enabling the computer or microprocessor to interact with a human user or other mechanism. Such software may include, but is not limited to, device drivers,

operating systems, and user applications. Additionally, such computer readable media may further include software for performing an embodiment of the present invention, as described above.

[0177] Included in the programming (software) of the general/specialized computer or microprocessor may be software modules for implementing the teachings of the present invention, including, but not limited to, inserting anchors into work artifacts, communication with application programming interfaces of various applications, initiating communications and communication clients, maintaining relative positions of conversation or communication clients to corresponding anchors in a work artifact, retrieving and logging conversations, requesting and handling communications requests, managing connections, initiating applications and downloading artifacts, and the display, storage, or communication of results according to embodiments of the present invention.

[0178] It should be understood that the particular embodiments described above are only illustrative of the principles of the present invention, and various modifications could be made by those skilled in the art without departing from the scope and spirit of the invention. Thus, the scope of the present invention is limited only by the claims that follow.